UTILIZING MOTION IN HUMANOID ROBOTS TO ENHANCE SPATIAL INFORMATION RECORDED BY MICROPHONE ARRAYS

Vladimir Tourbabin and Boaz Rafaely

Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev,
Beer-Sheva 84105, Israel, {tourbabv,br}@ee.bgu.ac.il

ABSTRACT

A recent and fast evolving application for microphone arrays is the auditory systems of humanoid robots. These arrays, in contrast to conventional arrays, are not fixed in a given position, but move together with the robot. While imposing a challenge to most conventional array processing algorithms, this movement offers an opportunity to enhance performance if utilized in an appropriate manner. The array movement can increase the amount of information gathered and, therefore, improve various aspects of array processing. This paper presents a theoretical framework for the processing of moving microphone arrays for humanoid robot audition based on a representation of the surrounding sound field in the spherical harmonics domain. A simulation study is provided, illustrating the use and the potential advantage of the proposed framework.

Index Terms— Microphone array, moving array, robot audition, rotation, translation, spherical harmonics.

1. INTRODUCTION

Microphone arrays of various configurations are being widely studied and have been successfully applied in various fields of modern engineering [1, 2]. One of the fast evolving applications of these arrays, which have received attention in recent years, is the auditory system of human-like robots or humanoids. Several publications describe humanoid robots capable of sound localization [3] and blind source separation [4] using microphone arrays. However, these arrays, in contrast to the convention, are not fixed at a given position, but move in accordance with the activity of the robot. Therefore, direct application of common processing methods is limited in this case.

Processing for the moving microphone array has been studied and described in recently published literature. For example, it was shown that it is possible to reduce spatial aliasing using a planar rotating array [5] or a linear array accelerating on a straight line [6]. In [7], a constant-speed motion was utilized in order to virtually decrease spacing between the sensors in a linear array moving along a straight line. In addition, the well-studied synthetic aperture array (SAR) technique utilizes the array motion for virtual extension of its aperture, and is widely applied in towed arrays in underwater acoustics [8, 9]. However, to the knowledge of the authors, there are no publications concerned with the processing of moving arrays for humanoid robot audition, which is the subject of this work.

The proposed approach models the effect of changing array positions on the measurements made by the array. The effect is modeled in the spherical harmonics (SH) domain, assuming that the microphones are distributed on the surface of the robot head [3, 4], which typically has a sphere-like shape. Using this model, the array measurements taken at different array positions are combined and can be used to virtually increase the number of measurement points and the array aperture. This approach can then be utilized to increase the number of acoustical sources that can be separated, improve the array directivity and increase the DOA estimation resolution as compared to stationary arrays. Furthermore, the proposed framework can be utilized for the development of active sensing, i.e. controlling the array motion in a way that maximizes the acquired information.

Sections 2 and 3 are devoted to describing the array model and its extension to account for array motion, followed, in 4, by a simulation investigating array performance with different modes of motion.

2. STATIONARY ARRAY MODEL

Consider an array of M microphones distributed on the surface of the head of a humanoid robot. Denote the i^{th} microphone output at a time t by $p_i(t)$. Denote the Fourier transform of the microphone output in a selected time window $0 \leq t < T$ by $P_i(\omega) = \mathcal{F}\{p_i(t), 0 \leq t < T\}$, where ω denotes the angular frequency. Next, assume that the surrounding sound field can be represented by a plane wave density function $a(\omega, \Omega)$, such that:

$$P_i(\omega) = \int_{\Omega \in S^2} v_i^*(\omega, \Omega) a(\omega, \Omega) d\Omega, \tag{1}$$

where $\Omega = (\theta, \phi)$ is a short notation for elevation θ and azimuth ϕ in the standard spherical coordinate system [10] with its origin located at the center of the head. Integral $\int_{\Omega \in S^2} d\Omega$ covers the entire surface of the unit sphere, denoted by S^2,

978-1-4799-3109-5/14/$31.00 ©2014 IEEE

$v_i^*(\omega, \Omega)$ is the i^{th} component of the array steering vector [11] and $(\cdot)^*$ denotes the complex conjugate operator. In the following, the dependence on ω will be omitted for clarity when possible.

The complex pressure amplitude measured by the i^{th} microphone can be rewritten as [10]:

$$P_i = \sum_{n=0}^{N} \sum_{m=-n}^{n} [v_{nm}^i]^* a_{nm}, \quad (2)$$

where a_{nm} and $v_{nm}^i(\Omega_i)$ are the spherical Fourier coefficients of the plane wave density function $a(\Omega)$ and of the complex conjugate of the i^{th} component of the array steering vector $v_i(\Omega)$, respectively. In (2) it is assumed that the SH order of the pressure amplitude on the head surface is limited to N. The expression for the maximum order N could be difficult to derive for a general head geometry. However, assuming that the head surface geometry is close to spherical, it is suggested here to use the expression for the effective order of the pressure on a rigid sphere, $N = \lceil kr \rceil$ [12], where $\lceil \cdot \rceil$ denotes the ceiling operator, r is the sphere radius and $k = \omega/c$, with c representing the speed of sound. Note that (2) also implies that in order to describe P_i the first N orders of coefficients a_{nm} are sufficient.

The steering vector component $v_i^*(\Omega)$ is also order-limited to N, implying that for sufficiently large Q

$$[v_{nm}^i]^* = \frac{4\pi}{Q} \sum_{q=1}^{Q} v_i^*(\Omega_q) Y_n^m(\Omega_q), \quad (3)$$

where it is assumed that $\{\Omega_q\}$ are nearly-uniformly distributed on the surface of a unit sphere (see [12] for details) with $Y_n^m(\cdot)$ denoting the SH of order n and degree m. Substituting (3) into (2) and rearranging leads to:

$$\begin{aligned} P_i &= \frac{4\pi}{Q} \sum_{q=1}^{Q} v_i^*(\Omega_q) \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_n^m(\Omega_q) a_{nm} \\ &= \mathbf{v}_i^H \mathbf{Y} \mathbf{a}_{nm}, \end{aligned} \quad (4)$$

where $\mathbf{a}_{nm} = [a_{0,0}\ a_{1,-1}\ a_{1,0}\ a_{1,1}\ \cdots\ a_{NN}]^T$, vector $\mathbf{v}_i^H = [v_i^*(\Omega_1)\ \cdots\ v_i^*(\Omega_Q)]$, $\mathbf{Y}$ is the inverse spherical Fourier transform matrix, whose entries are the spherical harmonics normalized by $4\pi/Q$, as suggested by (4), and operators $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and conjugate transpose operators, respectively. Concatenating the microphone outputs in a single measurement vector $\mathbf{P} = [P_1\ P_2\ \cdots\ P_M]^T$ and using (4), yields:

$$\mathbf{P} = \mathbf{V}\mathbf{Y}\mathbf{a}_{nm}, \quad (5)$$

where $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_M]^H$ with columns representing the steering vectors of the array.

The model in (5) can be used for the estimation of the plane wave density of the sound field, represented by $\mathbf{a}_{nm}$, from the array measurements taken at a single fixed array position. In the following section, this model is extended to account for arrays moving along an arbitrary trajectory.

3. MOVING ARRAY MODEL

In this section it is assumed that the array is moving. Suppose that the array motion is divided into L subsequent time windows, as illustrated in Fig. 1. Denote the Fourier transform of

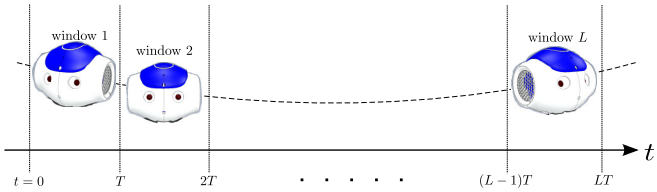


Fig. 1. Illustration of array motion divided into time windows.

the i^{th} microphone output in the l^{th} time window by:

$$P_i^l(\omega) = \mathcal{F}\{p_i(t),\ (l-1)T \le t < lT\},\ l = 1, 2, ..., L. \quad (6)$$

It is assumed here that the array velocity and the acceleration are sufficiently small, such that the effect of motion (e.g. Doppler shift) and the effect of varying array position within each time window can be neglected [6]. In this case, the measurements $\{P_i^l(\omega)\}_{i=1}^M$ obtained by the array in different time windows indexed by l differ by:

1. the time at which they were taken,
2. array position/orientation at which they were taken.

In order to account for the difference in time, assume that the sources producing the field are space stationary within the measurement time interval. Given that the sound field component at frequency ω is harmonic, the measurements taken at time window l can be aligned in time with respect to a common time frame, $l = 1$, by adding the appropriate phase shift:

$$\hat{P}_i^l(\omega) = P_i^l(\omega) \cdot e^{-j\omega(l-1)T},\ l = 1, 2, ..., L, \quad (7)$$

where $j = \sqrt{-1}$. The time aligned measurement vector in time window l is defined as:

$$\hat{\mathbf{P}}_l = [\hat{P}_1^l\ \hat{P}_2^l\ \cdots\ \hat{P}_M^l]^T, \quad (8)$$

where the dependence on frequency is omitted for notation simplicity. Now, using the model in (5), the measurement model for the time aligned l^{th} time window is given by:

$$\hat{\mathbf{P}}_l = \mathbf{V}\mathbf{Y}\mathbf{a}_{nm}^l, \quad (9)$$

where $\mathbf{a}_{nm}^l$ holds the SH coefficients of the plane wave density function at time window l relative to a coordinate system that is fixed with the array and the robot head. Thus, vectors $\mathbf{a}_{nm}^l$, $l = 1, 2, ..., L$ computed from (9) after time alignment, represent the sound field measured at the same time, but with microphones at different positions due to the motion of the array. The effect of changing array position is described in the following subsection.

3.1. Array rotation and translation

By taking the sound field relative to the array at $l = 1$ as a reference, the sound field relative to the array at time window l can be expressed as:

$$\mathbf{a}_{nm}^{l} = \mathbf{W}_l \mathbf{a}_{nm}^{1}, \tag{10}$$

where matrix $\mathbf{W}_l$ describes the effect of the spatial transformation of the sound field. Note that the spatial transformation described by $\mathbf{W}_l$ is reciprocal to the movement of the array relative to the field.

In order to specify $\mathbf{W}_l$ for a given movement, it is separated into two parts: a rotation about the origin of the coordinate system and a subsequent translation. This implies that $\mathbf{W}_l$ can be presented as:

$$\mathbf{W}_l = \mathbf{T}_l \mathbf{R}_l, \tag{11}$$

where $\mathbf{T}_l$ and $\mathbf{R}_l$ describe the translation and the rotation parts of the transformation, respectively.

The rotation matrix $\mathbf{R}_l$, is given by the following block diagonal matrix [13]:

$$\mathbf{R}_l = \begin{pmatrix} \mathbf{D}_0 & \mathbf{0}_{0,1} & \cdots & \mathbf{0}_{0,N} \\ \mathbf{0}_{1,0} & \mathbf{D}_1 & \cdots & \mathbf{0}_{1,N} \\ \vdots & & \ddots & \vdots \\ \mathbf{0}_{N,0} & \mathbf{0}_{N,1} & \cdots & \mathbf{D}_N \end{pmatrix}, \tag{12}$$

where $\mathbf{0}_{m_1,m_2}$ is a zero matrix having $2m_1 + 1$ rows and $2m_2 + 1$ columns and $\mathbf{D}_n \in \mathbb{C}^{(2n+1)\times(2n+1)}$ is given by:

$$\mathbf{D}_n = \begin{pmatrix} D^n_{-n,-n} & \cdots & D^n_{-n,n} \\ \vdots & \ddots & \vdots \\ D^n_{n,-n} & \cdots & D^n_{n,n} \end{pmatrix}, \tag{13}$$

where $D^{m_3}_{m_1,m_2}$ is the short notation for the Wigner-D function [14] $D^{m_3}_{m_1,m_2}(\alpha_l, \beta_l, \gamma_l)$ with α_l, β_l and γ_l being the Euler angles [10] of the rotation in the l^{th} time window. Note that $\mathbf{R}_l \in \mathbb{C}^{(N+1)^2\times(N+1)^2}$ is a square matrix, implying that rotation does not affect the SH order of the field.

The translation matrix $\mathbf{T}_l$, which describes the effect of the translation part of the sound field movement in the l^{th} time window, is given by [15]:

$$\mathbf{T}_l = \begin{pmatrix} \mathbf{\Gamma}_{0,0} & \mathbf{\Gamma}_{0,1} & \cdots & \mathbf{\Gamma}_{0,N} \\ \mathbf{\Gamma}_{1,0} & \mathbf{\Gamma}_{1,1} & \cdots & \mathbf{\Gamma}_{1,N} \\ \vdots & & \ddots & \vdots \\ \mathbf{\Gamma}_{N',0} & \mathbf{\Gamma}_{N',1} & \cdots & \mathbf{\Gamma}_{N',N} \end{pmatrix}, \tag{14}$$

where

$$\mathbf{\Gamma}_{n',n} = \begin{pmatrix} \gamma_{n',-n',n,-n} & \cdots & \gamma_{n',-n',n,n} \\ \vdots & \ddots & \vdots \\ \gamma_{n',n',n,-n} & \cdots & \gamma_{n',n',n,n} \end{pmatrix}, \tag{15}$$

with

$$\gamma_{n',m',n,m} = \sum_{q=0}^{\lceil kr_l \rceil} j_q(kr_l) \cdot Y_q^{m-m'}(\theta_l, \phi_l) \cdot C^{n,m,q}_{n',m'}, \tag{16}$$

where $j_q(\cdot)$ is the spherical Bessel function of order q and $Y_q^{m-m'}(\cdot)$ is the spherical harmonic function, as explained above, r_l and (θ_l, ϕ_l) are the distance and the direction of the translation in the l^{th} time window and $C^{n,m,q}_{n',m'}$ is a coefficient that involves the Wigner 3j symbols, as described in [15] (see equations (19) and (20) therein). The sum in (16) is limited to $\lceil kr_l \rceil$ because $j_q(kr_l) \approx 0$ for $q >> kr_l$ [12]. Note that the translation matrix $\mathbf{T}_l \in \mathbb{C}^{(N'+1)^2\times(N+1)^2}$ is, in general, not a square matrix; it transforms a field of order N into a field of order N', which, using the property that $C^{n,m,q}_{n',m'} = 0$ for $|n - n'| > q$, is given by

$$N' = N + \lceil kr_l \rceil. \tag{17}$$

This implies that a translation can effectively increase the SH order of the sound field by up to kr_l orders. Further details on the translation matrix can be found in [15] and references therein.

3.2. Combining measurements from various time windows

Eq. (9) relates the time-aligned array measurements in time window l to the sound field coefficients $\mathbf{a}^l_{nm}$, as viewed by the array. By using the representation introduced in (10) and (11) for rotation and translation, the measurements in different time windows indexed by l can be related to the same sound field, as viewed by the array in the (arbitrarily chosen) first time window:

$$\hat{\mathbf{P}}_l = \mathbf{V}\mathbf{Y}\mathbf{T}_l\mathbf{R}_l\mathbf{a}^1_{nm}, \quad l = 1, 2, ..., L. \tag{18}$$

Finally, all the measurements $\left\{\hat{\mathbf{P}}_l\right\}_{l=1}^{L}$ in (18) can be combined in a single measurement model by a column concatenation:

$$\hat{\mathbf{P}} = \mathbf{A}\mathbf{a}^1_{nm}, \tag{19}$$

where $\hat{\mathbf{P}} = [\hat{\mathbf{P}}_1^T\ \hat{\mathbf{P}}_2^T\ \cdots\ \hat{\mathbf{P}}_L^T]^T$ and

$$\mathbf{A} = \begin{pmatrix} \mathbf{V}\mathbf{Y}\mathbf{T}_1\mathbf{R}_1 \\ \mathbf{V}\mathbf{Y}\mathbf{T}_2\mathbf{R}_2 \\ \vdots \\ \mathbf{V}\mathbf{Y}\mathbf{T}_L\mathbf{R}_L \end{pmatrix}. \tag{20}$$

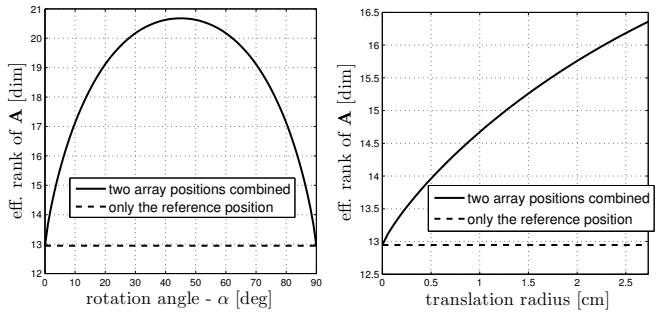
The model in (19) enables estimation of the sound field, represented by $(N + 1)^2$ SH coefficients of its plane wave density function, using all of the $M \cdot L$ measurements taken by M microphones in the subsequent L time windows, provided that the sound field does not change substantially in the time interval in which the measurements are taken.

The added value of combining measurements from different time windows may depend on the array trajectory, as further investigated in the numerical study below.

4. SIMULATION STUDY

Consider an acoustically rigid equiangular spherical array [12] of 13 microphones distributed with a $45°$ spacing in elevation and a $90°$ spacing in azimuth. The array radius is $r_a = 10$ cm and the SH order of the field is assumed to be limited to $N = 4$. In order to calculate $\mathbf{A}$ for a given array geometry, an expression for product $\mathbf{VY}$ of the steering matrix $\mathbf{V}$ and the inverse SFT matrix $\mathbf{Y}$ are required (see (20)). For the rigid spherical array this expression can be found in [12].

In the first simulation, the potential improvement in array performance as a result of a single displaced measurement position in addition to the reference position was examined. Two different modes of motion were considered: (a) rotation about the axis defined by $(\theta, \phi) = (0°, 0°)$ by angle α and (b) translation in the direction $(90°, 90°)$. The potential improvement in the array performance was quantified using the effective rank [16] of the transfer matrix $\mathbf{A}$ (see (17)). This measure has been previously shown to be related to the beamforming and DOA estimation performance of microphone arrays [17]. The effective rank of $\mathbf{A}$ at 2 kHz as a function of (a) the rotation angle α and (b) the distance of translation are presented in Fig. 2. It can be seen that the improvement, expressed by the



(a) (b)

Fig. 2. Effective rank of the combined transfer matrix $\mathbf{A}$ at 2 kHz (see (17)) as a function of (a) rotation angle, (b) translation distance.

effective rank of $\mathbf{A}$, increases for larger displacements (either rotation or translation). This is consistent with the fact that larger array displacement increases the effective array aperture and therefore can increase the spatial information gathered by the array about the surrounding sound field. In the rotation case, the improvement is maximal at $\alpha = 45°$. This is believed to be due to a spatial symmetry of the rotations by α and $90° - \alpha$ for this particular array geometry.

Another simulation was carried out in order to study the improvement induced by array movement as a function of frequency. Similar to the previous simulation, a single displaced measurement position was added to the reference position measurement. Three movements were considered: (a) rotation by $\alpha = 3°$, (b) translation by 5 mm and (c) a combination of both, with the rotation followed by the translation. The translation and rotation parameters were chosen to ensure that the translation distance and the displacement of the equator microphones due to the rotation by $3°$, are nearly the same. The effective rank of $\mathbf{A}$ as a function of frequency is plotted in Fig. 3.

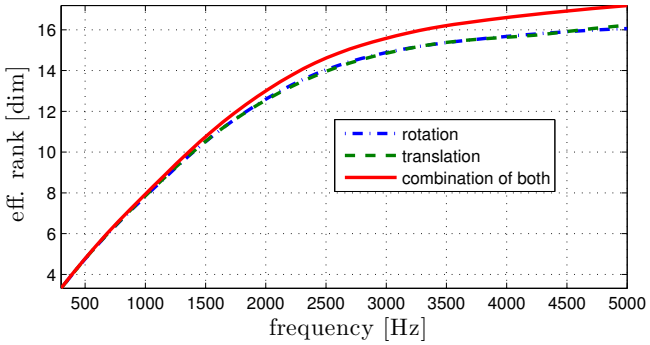


Fig. 3. Comparison of the effective rank of $\mathbf{A}$ as a function of frequency for three different types of movements.

It can be seen that in all three cases the improvement increases with frequency. Note that the improvements in the effective rank due to the rotation or the translation alone are nearly identical. This implies that the effect of rotation and translation on the effective rank can be similar. In addition, the effect of the combined movement (rotation followed by translation) is higher than that of the rotation or the translation alone.

5. CONCLUSION

A theoretical framework for the processing of moving microphone arrays for humanoid robots was proposed. This framework has the potential to improve various aspects of array processing by combining the spatial information acquired by the array at different spatial positions. Simulation examples showed that the performance improvement might depend on the array trajectory and, in general, will increase for higher array displacements. Future work is expected to include an investigation of the effect of the stationarity of the sound field and an evaluation of the possible decrease in performance due to the self-noise produced by a moving robot.

6. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465.

7. REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer topics in signal processing. Springer, 2008.

[2] I. Cohen, J. Benesty, and S. Gannot, *Speech Processing in Modern Communication: Challenges and Perspectives*, Springer topics in signal processing. Springer-Verlag, 2009.

[3] N. Schmitz, C. Spranger, and K. Berns, "3D audio perception system for humanoid robots," in *ACHI 2009*, Feb. 2009, pp. 181–186.

[4] M. Maazaoui, K. Abed-Meraim, and Y. Grenier, "Adaptive blind source separation with HRTFs beamforming preprocessing," in *SAM 2012*, June 2012, pp. 269–272.

[5] A. Cigada, M. Lurati, F. Ripamonti, and M. Vanali, "Moving microphone arrays to reduce spatial aliasing in the beamforming technique: Theoretical background and numerical investigation," *J. Acoust. Soc. Am.*, vol. 124, no. 6, pp. 3648–3658, 2008.

[6] E. Chang, "Irregular array motion and extended integration for the suppression of spatial aliasing in passive sonar," *J. Acoust. Soc. Am.*, vol. 129, no. 2, pp. 765–773, 2011.

[7] M. H. Johnson, "Synthetic elements for moving line arrays," in *OCEANS 2009 - EUROPE*, 2009, pp. 1–6.

[8] J. A. Fawcett, "Synthetic aperture processing for a towed array and a moving source," *J. Acoust. Soc. Am.*, vol. 94, no. 5, pp. 2832–2837, 1993.

[9] S. Kim, D. H. Youn, and C. Lee, "Temporal domain processing for a synthetic aperture array," *IEEE J. Ocean. Eng.,*, vol. 27, no. 2, pp. 322–327, Apr 2002.

[10] G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists*, Mathematical Methods for Physicists. Elsevier, 2005.

[11] H. L. Van Trees, *Optimum Array Processing (Detection, Estimation and Modulation Theory, Part IV)*, Wiley Interscience, New York, 2002.

[12] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, 2005.

[13] P. J. Kostelec and D. N. Rockmore, "FFTs on the rotation group," *J. Fourier Anal. and Appl.*, vol. 14, no. 2, pp. 145–179, 2008.

[14] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskiĭ, *Quantum Theory of Angular Momentum*, World Scientific Publishing Company, Incorporated, 1988.

[15] T. Peleg and B. Rafaely, "Investigation of spherical loudspeaker arrays for local active control of sound," *J. Acoust. Soc. Am.*, vol. 130, no. 4, pp. 1926–1935, 2011.

[16] O. Roy and M. Vetterli, "The effective rank: A measure of effective dimentionality," in *EUSIPCO*, Sep. 2007, pp. 606–610.

[17] V. Tourbabin and B. Rafaely, "Theoretical framework for the design of microphone arrays for robot audition," in *ICASSP, IEEE*, 2013, pp. 4290–4294.